Journal of Data Science 24 (4), 646–665
October 2026
DOI: 10.6339/25-JDS1209
Statistical Data Science

Rating Competitors in Games with Strength-Dependent Tie Probabilities

Mark E. Glickman*
[1]*Department of Statistics, Harvard University, Cambridge, MA, USA*

Abstract

Competitor rating systems for head-to-head games are typically used to measure playing strength from game outcomes. Ratings computed from these systems are often used to select top competitors for elite events, for pairing players of similar strength in online gaming, and for players to track their own strength over time. Most implemented rating systems assume only win/loss outcomes, and treat occurrences of ties as the equivalent to half a win and half a loss. However, in games such as chess, the probability of a tie (draw) is demonstrably higher for stronger players than for weaker players, so that rating systems ignoring this aspect of game results may produce strength estimates that are unreliable. We develop a new rating system for head-to-head games based on a model that explicitly acknowledges that a tie may depend on the strengths of the competitors. The approach uses a Bayesian dynamic modeling framework. Within each time period, posterior updates are computed in closed form using a single Newton-Raphson iteration evaluated at the prior mean. The approach is demonstrated on a large dataset of chess games played in International Correspondence Chess Federation tournaments.

Keywords *Bayesian dynamic generalized linear model; Bradley-Terry model; chess tournaments; order effects; Paired comparison models; ranking models; tie outcomes*

1 Introduction

Rating systems for chess play a vital role in organizing tournaments and matches, tracking individual progress, and facilitating fair matchmaking. They provide a standardized framework to compare the strength of players based on game outcomes, enabling national and international federations to assign titles, seed tournaments, and monitor competitive balance across player pools (Glickman, 1995). As competitive chess continues to integrate into global and online formats, the need for dynamic, data-driven rating systems has never been greater.

Numerous systems have been proposed to address the need for reliable, scalable rating methods in chess and other competitive settings. Over the past decades, several chess rating systems have gained widespread adoption. Among the most influential is the system developed by Elo (1978), introduced in the late 1950s and implemented by the US Chess Federation in the early 1960s. It was subsequently adopted by the World Chess Federation (FIDE) in the 1970s. Elo's approach is loosely based on the model of Bradley and Terry (1952), a foundational probability model for paired comparisons. Building on this framework, the Glicko (Glickman, 1999) and Glicko-2 (Glickman, 2001) systems were introduced in the 1990s and early 2000s. These models approximate Bayesian state-space frameworks in which game outcomes are governed by

* Email: glickman@fas.harvard.edu.

© 2026 The Author(s). Published by the School of Statistics and the Center for Applied Statistics, Renmin University of China. Open access article under the CC BY license.
Received June 13, 2025; Accepted November 25, 2025

the Bradley-Terry model and player abilities evolve according to Gaussian processes. Today, Elo-based and Glicko-style systems are widely used by national and international organizations, online platforms, and competitive gaming leagues, serving as essential tools for assessing player strength in large-scale environments. A comprehensive review of rating systems currently in use can be found in Glickman and Jones (2024).

The International Correspondence Chess Federation (ICCF), the principal international body for organizing competitive correspondence chess, has long relied on a rating system based largely on Elo's formulation. However, the system has shown critical shortcomings when applied to the distinctive features of ICCF competition. Chief among these is the exceptionally high rate of drawn games among top-level players. The Elo, Glicko, and Glicko-2 systems do not explicitly model the probability of a draw; instead, they treat a draw as equivalent in informational value to the average of a win and a loss. While Szczecinski and Djebbi (2020) introduced a modification of the Elo framework, building on the model by Davidson (1970), to incorporate draws explicitly, this extension assumes a constant draw frequency and does not allow the probability of a draw to vary with player strength. In ICCF play, approximately 95% of games between top-rated players end in a draw, whereas lower-rated players draw far less frequently. This strength-dependent draw behavior is not adequately captured by rating systems that lack a mechanism for modeling tie probabilities as a function of ability. As a result, these systems may produce stagnant ratings, especially for elite players whose results are overwhelmingly drawn. Recognizing these structural limitations, ICCF officials initiated efforts to reevaluate their rating methodology and explore alternative systems better suited to the unique dynamics of correspondence chess.

This paper introduces a new rating system designed to address limitations of existing methods in modeling draw behavior. The proposed system builds on the recently developed model by Glickman (2025), which explicitly models the probability of a draw as a function of player strength. In addition, it incorporates a dynamic framework that allows player ratings to evolve over time. Player abilities are modeled as following a normal random walk, capturing natural fluctuations in strength, and a Bayesian filtering procedure is used to update ratings based on recent game outcomes. The system balances the need to incorporate historical performance with the practical goal of emphasizing current ability. By more accurately modeling draw tendencies and supporting time-sensitive updates, the method provides a principled and scalable solution for rating players in correspondence chess and related competitive settings. The ICCF officially adopted and implemented this rating system in 2023.

This paper is organized as follows. Section 2 provides background on the challenges faced by the ICCF, including an empirical demonstration of the limitations in their implementation of the Elo rating system. In Section 3, we introduce a fully Bayesian state-space model for game outcomes, building on the framework developed in Glickman (2025). The model accounts for strength-dependent draw probabilities, incorporates the possibility that stronger players may better exploit the advantage of playing white, and allows player abilities to evolve over time. In Section 4, we present an approximate filtering algorithm that translates a one-step Bayesian updating process into a computationally efficient method for real-time rating updates. Section 6 applies the proposed approach to ICCF game outcomes from 2016 to 2022. We conclude in Section 7 with a discussion of findings and implications.

2 ICCF Background

The ICCF, recognized by FIDE as the official global authority for correspondence chess, was established in 1951. It oversees international correspondence competitions conducted via postal mail and, more recently, through an online server platform. The ICCF maintains a structured system of player titles and ratings, closely aligned with those used in over-the-board play. Unlike traditional chess tournaments, correspondence games unfold over extended periods, often allowing weeks or even months per move. Players are permitted to use computational resources such as databases, opening references, and chess engines, although engine use is regulated in certain events. The ICCF organizes both individual and team championships, including world title events. With the transition from postal to digital formats in recent decades, the ICCF now facilitates play primarily through its online platform, enabling thousands of players worldwide to engage in structured, asynchronous international competition.

In 2021, I was invited by the ICCF to develop a new rating system to replace its long-standing Elo-based framework. The Elo (1978) system updates a player's rating, an estimate of their playing strength, based on game outcomes using relatively simple computations. A player's rating increases after a draw against a higher-rated opponent or after a win, and decreases after a draw against a lower-rated opponent or after a loss. Elo ratings are scaled so that a rating of 1500 corresponds to average strength, ratings of 1000 or below indicate relatively weak players, and ratings above 2000 correspond to expert-level strength. Ratings of 2500 and higher are typically associated with world-class players. If R_i and R_j denote the Elo ratings of players i and j, respectively, the expected score, or "winning expectancy" in Elo's terminology, for player i is given by the formula

$$\frac{1}{1+10^{-(R_i-R_j)/400}},$$

which is a rescaled inverse-logit function of the rating difference.

The ICCF's initiative was motivated by well-documented concerns about the limitations of its existing rating system, particularly its inability to accommodate the distinctive features of correspondence chess. Foremost among these concerns was the extremely high frequency of drawn games at the highest levels of play. A primary factor contributing to this pattern is the widespread use of powerful chess engines by top players to support their analysis. Given the current capabilities of these engines, it is exceedingly difficult for even the strongest players to secure a meaningful advantage when both sides are making near-optimal moves. Consequently, the Elo-based system struggled to register informative rating changes from drawn outcomes and failed to account for how the value of a draw might vary with the relative strength of the opponents. This limitation resulted in rating stagnation and reduced differentiation among elite players, underscoring the need for a new system better suited to the realities of modern correspondence chess.

The relationship between draw frequency and player strength in ICCF competition is evident from an empirical analysis. We examined all ICCF games completed between 2016 and 2022 in which the pre-game rating difference between the two players was less than 50 points. These pre-game ICCF ratings were used as proxies for player strength, acknowledging that the broader objective of this work is to develop a more refined rating algorithm. A 50-point difference corresponds to a winning expectancy of approximately 0.57 for the higher-rated player, so that rating differences less than 50 reflect matchups between players of similar ability. This restriction yielded a dataset of 149,731 game outcomes. To explore the relationship between draw probability and player strength, we fit a logistic regression model for the probability that a

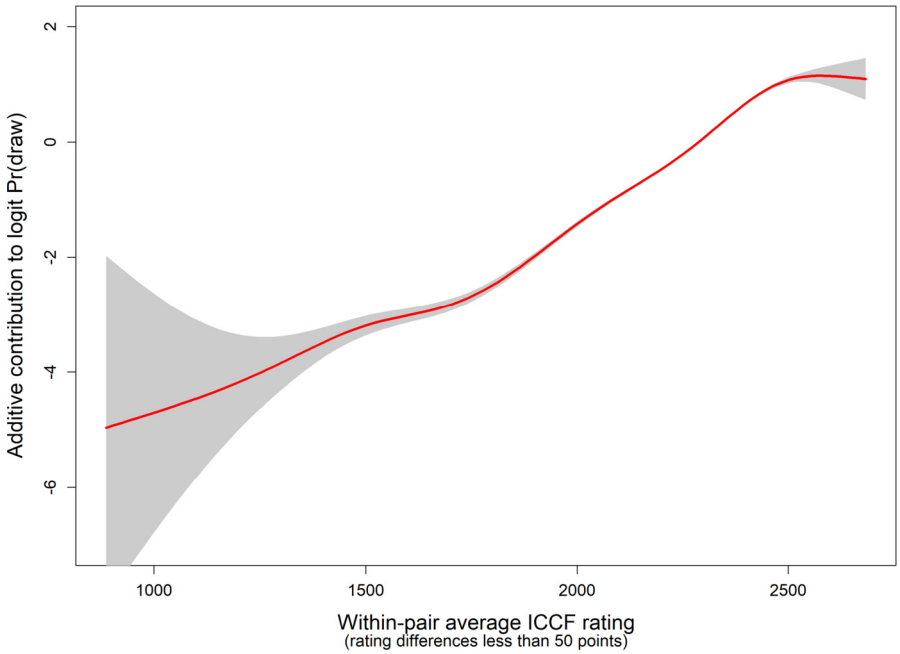

Figure 1: Log-odds contribution of drawing as a function of the average rating between two players, controlling for absolute rating difference. The shaded area represents 95% pointwise confidence limits.

game resulted in a draw (as opposed to a decisive outcome), using generalized additive modeling (Hastie and Tibshirani, 1986; Wood, 2011) implemented in the `mgcv` package in R. The model included smoothing spline terms for both the absolute rating difference and the average rating of the two players.

Figure 1 illustrates the relationship between the average ICCF rating of two players and the additive contribution to the log-odds of a game ending in a draw. The red curve represents the estimated smooth effect, while the shaded region indicates a 95% pointwise confidence band. The plot shows a strong positive linear association: as the average player strength increases, the log-odds of a draw also increases. This implies that higher-rated players, those more likely to use strong engine-assisted analysis, are significantly more prone to drawing games. The effect appears to level off slightly at the highest rating levels, suggesting a saturation point where draws are nearly inevitable.

3 A Bayesian State-Space Model for Paired Comparisons

The development of a revised rating system for the ICCF begins with the specification of a probability model for game outcomes. We adopt a Bayesian framework, consistent with the foundations of the Glicko and Glicko-2 rating systems. Our starting point is the model proposed by Glickman (2025), which explicitly accounts for the probability of a draw as a function of player strength.

Consider M chess players in a league, and divide time into T equally spaced periods. Let $\theta_{1t}, \ldots, \theta_{Mt}$ denote the latent strength parameters of the players at time period t, for $t = 1, \ldots, T$.

For a game between players i and j in period t, define the game outcome variable as

$$Y_{ijt} = \begin{cases} 0 & \text{if player } i \text{ loses to opponent } j \text{ in period } t, \\ 1/2 & \text{if player } i \text{ ties/draws opponent } j \text{ in period } t, \\ 1 & \text{if player } i \text{ defeats opponent } j \text{ in period } t. \end{cases}$$

Following Glickman (2025), we model the probabilities of each possible outcome as

$$\mathrm{P}(Y_{ijt} = y) = \begin{cases} \dfrac{\exp\left(\theta_{it} + x_{ijt}(\alpha_0 + \alpha_1 \cdot \frac{\theta_{it}+\theta_{jt}}{2})/4\right)}{S_{ijt}} & \text{if } y = 1, \\ \dfrac{\exp\left(\theta_{jt} - x_{ijt}(\alpha_0 + \alpha_1 \cdot \frac{\theta_{it}+\theta_{jt}}{2})/4\right)}{S_{ijt}} & \text{if } y = 0, \\ \dfrac{\exp\left(\beta_0 + (1+\beta_1) \cdot \frac{\theta_{it}+\theta_{jt}}{2}\right)}{S_{ijt}} & \text{if } y = 1/2, \end{cases} \tag{3.1}$$

where $x_{ij} = 1$ if player i has the white pieces and -1 if playing black, and where α_0, α_1, β_0 and β_1 are real-valued parameters shared across all player pairs. The normalizing constant S_{ijt} is the sum of the numerators,

$$\begin{aligned} S_{ijt} \;=\; & \exp\left(\theta_{it} + x_{ijt}\left(\alpha_0 + \alpha_1 \cdot \frac{\theta_{it}+\theta_{jt}}{2}\right)/4\right) + \exp\left(\theta_{jt} - x_{ijt}\left(\alpha_0 + \alpha_1 \cdot \frac{\theta_{it}+\theta_{jt}}{2}\right)/4\right) \\ & + \exp\left(\beta_0 + (1+\beta_1) \cdot \frac{\theta_{it}+\theta_{jt}}{2}\right). \end{aligned}$$

Several observations about the model in (3.1) are worth highlighting. When $\alpha_1 = \beta_1 = 0$, the model reduces to a non-dynamic form proposed by David (1988), in which the log-odds of winning, ignoring ties, becomes $\log(\mathrm{P}(Y_{ijt} = 1)/\mathrm{P}(Y_{ijt} = 0)) = \theta_{it} - \theta_{jt} + (x_{ijt}/2)\alpha_0$. This is the no-tie model by Davidson and Beaver (1977), a special case of David (1988), which justifies the division by 4 in the exponents of (3.1). In the case of $\alpha_1 = \beta_1 = 0$, the parameter α_0 captures a fixed advantage for the player with the white pieces: a positive value indicates an advantage for white, while a negative value indicates an advantage for black. The parameter β_0 governs the overall frequency of draws, with larger values corresponding to higher draw probabilities. The model in (3.1) extends that of David (1988) by allowing both the draw probability and the white advantage to vary as functions of player strength. Specifically, when $\alpha_1 > 0$, the probability that white wins increases with the average strength of the two players, controlling for their strength difference. Similarly, when $\beta_1 > 0$, the probability of a draw increases with the average player strength. Further details on this model and its implications can be found in Glickman (2025).

To account for changes in player ability over time, we incorporate a dynamic component into the model via a stochastic innovation process. Specifically, each player's latent strength is assumed to evolve according to a normal random walk across discrete time periods. For $t = 2, \ldots, T$,

$$\theta_{i,t+1} \sim \mathrm{N}(\theta_{it}, \tau^2), \tag{3.2}$$

where τ^2 is the innovation variance governing the variability in strength between consecutive periods. This specification assumes nonstationarity in player strength, with variability accumulating over time. This formulation allows player abilities to fluctuate incrementally over time,

capturing natural changes due to factors such as practice, aging, or varying levels of competitive engagement. By embedding this stochastic structure, the model remains responsive to recent performance while preserving reasonable estimates for players with sparse or infrequent activity.

This modeling strategy aligns with prior work on time-varying strength estimation, including methods proposed by Fahrmeir and Tutz (1994), Glickman (1999), and Ingram (2021), which employ stochastic processes to capture longitudinal variation in latent ability. The normal random walk assumption introduces temporal smoothness, helping to avoid overfitting to short-term fluctuations while still capturing meaningful trends in player performance. Although autoregressive formulations could offer more structured dynamics and the advantage of a stationary distribution, they pose challenges, particularly when players have no observed outcomes in a given period, leading to parameter drift. These limitations make the normal random walk a more robust and practical choice for modeling evolving player strength over time.

The Bayesian modeling formulation is completed by specifying a prior distribution for the strength parameters at time $t = 1$. For each player i, the initial latent strength is assumed to follow a normal distribution with player-specific mean μ_{i1} and variance σ_{i1}^2,

$$\theta_{i1} \sim \mathrm{N}(\mu_{i1}, \sigma_{i1}^2), \tag{3.3}$$

for $i = 1, \ldots, M$. This prior encodes baseline beliefs about each player's ability before any game outcomes are observed. When historical data or prior ratings are available, such as from earlier rating systems or previous competitions, they can inform the choice of μ_{i1} and σ_{i1}^2, grounding the model in established assessments of skill. For experienced players with extensive performance histories, smaller values of σ_{i1}^2 reflect greater confidence in their prior ratings. In contrast, for new players with no prior games recorded, the prior parameters may be set using default values based on demographic characteristics or assumed to reflect the broader population distribution of playing strengths.

The combination of the three model components in (3.1), (3.2) and (3.3) defines a dynamic generalized linear model (West et al., 1985), a specific instance of a broader class of state-space models (Durbin and Koopman, 2012). The conditional posterior distribution of the strength parameters, given game outcomes $\boldsymbol{y}$ over all T time periods and the non-strength parameters, is given by

$$\begin{aligned} &p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_T \mid \alpha_0, \alpha_1, \beta_0, \beta_1, \tau^2, \boldsymbol{y}) \\ &\quad \propto \quad p(\boldsymbol{\theta}_1) \prod_{t=1}^{T} p(\boldsymbol{y}_t \mid \boldsymbol{\theta}_t, \alpha_0, \alpha_1, \beta_0, \beta_1) \prod_{t=2}^{T} p(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t-1}, \tau^2), \end{aligned}$$

where $\boldsymbol{y}_t$ denotes the vector of game results during period t, and $\boldsymbol{\theta}_t = (\theta_{1t}, \ldots, \theta_{Mt})$ is the vector of player strengths at time t. Treating the non-strength parameters as fixed and known, or by placing priors on them, models of this form have traditionally been analyzed via Markov chain Monte Carlo (MCMC) methods. Examples include the work of Glickman (1993), Glickman (1999), Knorr-Held (2000) and Gorgi et al. (2019). While MCMC-based approaches are effective in settings with a modest number of competitors and time periods, they become computationally infeasible as the scale of the data increases. In environments such as league play or online gaming platforms, where thousands of participants compete across many time intervals, the dimensionality of the parameter space renders full posterior simulation burdensome. Furthermore, in a real-time applications, the primary goal is often to estimate current player strength rather than reconstruct historical trajectories. This practical requirement calls for a more efficient approach

that can update ratings sequentially as new outcomes are observed, without reprocessing the entire history of play.

To address these challenges, we propose in Section 4 a one-step-ahead filtering algorithm that recursively approximates the posterior distributions of player strength using only the most recent game outcomes. This approach enables efficient, real-time updates without the need to reprocess the full history of results. Similar filtering strategies have been applied in simpler paired comparison contexts, as in Glickman (1999) and Glickman (2001).

4 Approximate Filtering Algorithm for Real-Time Rating Updates

Rather than analyze the full posterior distribution after all game outcomes in period t, we propose a one-step filtering recursion that sequentially updates inferences about player strengths, first based on game outcomes, and then based on the passage of time. Let $\boldsymbol{D}_t$ denote the collection of all game outcomes up to and including period t. At the start of period t, before game results are observed, we assume that for each player $i = 1, \ldots, M$, the mean μ_{it} and variance σ_{it}^2 of their strength distribution are known. We also assume, for now, that the parameters α_0, α_1, β_0, β_1 and τ^2, which we henceforth refer to as "system parameters," are fixed and known.

Under these assumptions, the prior distribution for θ_{it} is given by

$$\theta_{it} \mid \boldsymbol{D}_{t-1} \sim \mathrm{N}(\mu_{it}, \sigma_{it}^2). \tag{4.1}$$

To update this prior using the outcomes from period t, we apply a multinomial likelihood based on the probabilities defined in (3.1), yielding an approximate posterior distribution for period t

$$\theta_{it} \mid \boldsymbol{D}_t \sim \mathrm{N}(\mu_{it}^*, \sigma_{it}^{*2}), \tag{4.2}$$

where μ_{it}^* and σ_{it}^{*2} are the approximate posterior mean and variance, respectively.

To obtain the prior distribution for period $t+1$, we apply the law of total expectation and variance, using the dynamic model in (3.2). For normal densities, this yields

$$\begin{aligned} \mathrm{E}(\theta_{i,t+1} \mid \boldsymbol{D}_t) &= \mathrm{E}(\mathrm{E}(\theta_{i,t+1} \mid \theta_{it}, \boldsymbol{D}_t)) = \mathrm{E}(\theta_{it} \mid \boldsymbol{D}_t) = \mu_{it}^* \,. \\ \mathrm{Var}(\theta_{i,t+1} \mid \boldsymbol{D}_t) &= \mathrm{E}(\mathrm{Var}(\theta_{i,t+1} \mid \theta_{it}, \boldsymbol{D}_t)) + \mathrm{Var}(\mathrm{E}(\theta_{i,t+1} \mid \theta_{it}, \boldsymbol{D}_t)) \\ &= \mathrm{E}(\tau^2 \mid \boldsymbol{D}_t) + \mathrm{Var}(\theta_{it} \mid \boldsymbol{D}_t) = \tau^2 + \sigma_{it}^{*2}. \end{aligned}$$

Therefore, the prior distribution for period $t+1$ becomes

$$\theta_{i,t+1} \mid \boldsymbol{D}_t \sim \mathrm{N}(\mu_{it}^*, \sigma_{it}^{*2} + \tau^2), \tag{4.3}$$

which may then be used to process game results from period $t+1$, allowing the recursion to proceed forward through time.

It is important to note that, provided the posterior distribution of θ_{it} is approximately normal, the transition from the posterior in (4.2) to the prior in (4.3) represents an exact probabilistic update. Our overall strategy thus relies on being able to approximate the posterior density that arises from combining the normal prior in (4.1) with the multinomial likelihood. This approximation yields the normal form given in (4.2). In the sections that follow, we outline a procedure for computing the approximate posterior mean μ_{it}^* and variance σ_{it}^{*2} for each player i. We summarize below the main steps of the approximation.

- Section 4.1: To obtain a tractable expression for the posterior distribution of θ_{it} averaged over opponents' strength distributions, we make a conservative simplification by replacing opponents' posterior strength distributions with their prior distributions. This assumption yields an approximate marginal posterior density for θ_{it} that is conditionally independent across players, making the subsequent updates separable.
- Section 4.2: Because the marginal posterior density involves integrals with no closed-form solution, we approximate them using Gauss–Hermite quadrature over a small grid of nodes. This produces a closed-form expression for the marginal posterior density of θ_{it}.
- Section 4.3: Finally, the posterior mean and standard deviation are approximated from this marginal density using a one-step Newton–Raphson update of the log-marginal posterior, evaluated at the prior mean and variance.

4.1 Opponent Prior Approximation and Marginalization

To update the prior distribution of θ_{it} for player i based on game outcomes during period t, we begin by replacing the posterior distributions of player i's opponents with their corresponding prior distributions. Letting $-i$ denote the indices of all players other than i, we assume

$$p(\boldsymbol{\theta}_{-it} \mid \boldsymbol{D}_t) \approx p(\boldsymbol{\theta}_{-it} \mid \boldsymbol{D}_{t-1}).$$

This approximation is expected to be accurate when the opponents of player i have well-informed prior estimates of strength and play relatively few games during period t. Even when these conditions are not fully met, such as when some opponents have weaker priors or play many games during period t, the effect of using an opponent's prior rather than the posterior remains conservative. By treating opponents' strengths as slightly more uncertain than they would be under full updating, the approximation tends to produce posterior distributions for player i that are a bit more variable rather than overconfident. In this sense, the approximation errs on the side of caution. A similar approximation was employed in Glickman (1999) in the development of the Glicko rating system. With this approximation, the joint posterior distribution of $\boldsymbol{\theta}_t$ can be approximated as

$$\begin{aligned} p(\boldsymbol{\theta}_t \mid \boldsymbol{D}_t) &= p(\theta_{it} \mid \boldsymbol{\theta}_{-it}, \boldsymbol{D}_t) p(\boldsymbol{\theta}_{-it} \mid \boldsymbol{D}_t) \\ &\approx p(\theta_{it} \mid \boldsymbol{\theta}_{-it}, \boldsymbol{D}_t) p(\boldsymbol{\theta}_{-it} \mid \boldsymbol{D}_{t-1}) \\ &= p(\theta_{it} \mid \boldsymbol{\theta}_{-it}, \boldsymbol{D}_t) \prod_{j \neq i} \mathrm{N}(\theta_{jt} \mid \mu_{jt},\ \sigma_{jt}^2), \end{aligned} \tag{4.4}$$

where $\mathrm{N}(\cdot \mid \mu, \sigma^2)$ denotes a normal density function with mean μ and variance σ^2.

In (4.4), we recognize that $\boldsymbol{D}_t = (\boldsymbol{y}_t, \boldsymbol{D}_{t-1})$, allowing the first factor to be expanded using Bayes' rule,

$$\begin{aligned} p(\theta_{it} \mid \boldsymbol{\theta}_{-it}, \boldsymbol{D}_t) &\propto p(\theta_{it} \mid \boldsymbol{\theta}_{-it}, \boldsymbol{D}_{t-1}) p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{\theta}_t, \boldsymbol{D}_{t-1}) \\ &= \mathrm{N}(\theta_{it} \mid \mu_{it}, \sigma_{it}^2) p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{\theta}_t), \end{aligned} \tag{4.5}$$

with $\boldsymbol{x}_t$ representing the vector of x_{ijt} for player i during period t, and where the final equality uses the assumption that $\theta_{it} \mid \boldsymbol{\theta}_{-it}, \boldsymbol{D}_{t-1} \sim \mathrm{N}(\mu_{it}, \sigma_{it}^2)$, and that the likelihood depends only on the current period's game outcomes and player strengths.

Substituting into (4.4) and marginalizing over the opponent strengths $\boldsymbol{\theta}_{-it}$, we obtain

$$\begin{aligned} p(\theta_{it} \mid \boldsymbol{D}_t) &= \int p(\boldsymbol{\theta}_t \mid \boldsymbol{D}_t) d\boldsymbol{\theta}_{-it} \approx \int \left(p(\theta_{it} \mid \boldsymbol{\theta}_{-it}, \boldsymbol{D}_t) \prod_{j \neq i} \mathrm{N}(\theta_{jt} \mid \mu_{jt},\ \sigma_{jt}^2) \right) d\boldsymbol{\theta}_{-it} \\ &\propto \mathrm{N}(\theta_{it} \mid \mu_{it}, \sigma_{it}^2) \int \left(p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{\theta}_t) \prod_{j \neq i} \mathrm{N}(\theta_{jt} \mid \mu_{jt},\ \sigma_{jt}^2) \right) d\boldsymbol{\theta}_{-it} \\ &\propto \mathrm{N}(\theta_{it} \mid \mu_{it}, \sigma_{it}^2) \prod_{j \in \Omega_{it}} \int p(y_{ijt} \mid x_{ijt}, \theta_{it}, \theta_{jt}) \mathrm{N}(\theta_{jt} \mid \mu_{jt}, \sigma_{jt}^2) d\theta_{jt}, \end{aligned} \tag{4.6}$$

where Ω_{it} denotes the index set of opponents of player i during period t. The term $p(y_{ijt} \mid x_{ijt}, \theta_{it}, \theta_{jt})$ refers to the game outcome probability model defined in (3.1), expressed as a function of the strength parameters, and with dependence on the system parameters suppressed for clarity. The approximation in the first line of (4.6) follows from (4.4), and the proportionality from the first to the second line results from substituting the expression in (4.5).

Equation (4.6) assumes that each pair of players competes at most once during a given time period t. In practice, however, the same two players may face each other multiple times within the same rating period. To accommodate this, we treat each repeated game as if it were played against a distinct opponent having the same prior strength distribution. This approximation preserves the structural simplicity of the computational approach while allowing multiple game results against the same opponent to be incorporated. It maintains the overall informativeness of the data without introducing statistical dependencies that would complicate inference.

4.2 Gauss-Hermite Approximation of Multinomial Logit Integrals

The integrals in (4.6), which take the form of normal mixtures of multinomial logit probabilities (with non-strength parameters treated as known), do not admit straightforward analytic approximations. In settings where the likelihood is based on inverse-logit expressions, as in Glickman (1999), a range of approximation techniques have been developed for evaluating normal mixtures of logistic functions. These include methods proposed by Boys and Dunsmore (1987), Crouch and Spiegelman (1990), and Pirjol (2013). However, When the outcome probabilities follow a multinomial logit model, as in our case, the set of available analytic approximations is considerably more limited.

Several methods are available for approximating normal mixtures of multinomial logit probabilities. The most direct is to use MCMC simulation; however, such iterative techniques are computationally impractical in the context of rating large numbers of players, where efficiency is critical. As an alternative, closed-form or semi-analytic approximations have been proposed. For example, Harding and Hausman (2007) applied a multivariate Laplace approximation to evaluate the required integrals. More closely aligned with the approach we adopt below, Bhat (1995) and Pryanishnikov and Zigova (2016) approximate the normal mixture integrals using Gaussian quadrature methods, which offer a computationally efficient and numerically stable solution for problems of this structure.

To approximate the integrals in (4.6), we adopt a 2-point Gauss-Hermite quadrature strategy (Steen et al., 1969). This method provides a simple yet effective numerical approximation for integrals involving normal densities. In particular, we approximate the continuous integrals

in (4.6) by

$$\int p(y_{ijt} \mid x_{ijt}, \theta_{it}, \theta_{jt}) \mathrm{N}(\theta_{jt} \mid \mu_{jt}, \sigma_{jt}^2) d\theta_{jt} \approx$$
$$\frac{1}{2} p(y_{ijt} \mid x_{ijt}, \theta_{it}, \mu_{jt} - \sigma_{jt}) + \frac{1}{2} p(y_{ijt} \mid x_{ijt}, \theta_{it}, \mu_{jt} + \sigma_{jt}),$$

which corresponds to a symmetric 2-point approximation centered at the mean of the normal distribution, with evaluation points located one standard deviation above and below the mean. Using the Gauss-Hermite approximation, the expression in (4.6) becomes

$$p(\theta_{it} \mid \boldsymbol{D}_t) \propto \mathrm{N}(\theta_{it} \mid \mu_{it}, \sigma_{it}^2) \prod_{j \in \Omega_{it}} \left(\frac{1}{2} p(y_{ijt} \mid x_{ijt}, \theta_{it}, \mu_{jt} - \sigma_{jt}) + \frac{1}{2} p(y_{ijt} \mid x_{ijt}, \theta_{it}, \mu_{jt} + \sigma_{jt}) \right). \tag{4.7}$$

Although the approximate posterior for θ_{it} in (4.7) does not have a closed-form expression, it can be evaluated numerically and summarized through standard numerical methods. However, in the context of a rating system, where updates must be performed frequently and efficiently, it is important to retain computational tractability. To this end, we approximate the posterior with a normal distribution in closed form by constructing a Gaussian density that matches the mode and curvature of the log-posterior. This approach, detailed in the following section, yields a surrogate posterior distribution that enables player strengths to be updated in a simple, scalable, and analytically convenient manner.

4.3 Closed-Form Normal Approximation via Newton-Raphson

The approximating log-posterior density for θ_{it}, based on (4.7), can be written as

$$\begin{aligned} \log p(\theta_{it} \mid \boldsymbol{D}_t) = & \\ & c + \log \mathrm{N}(\theta_{it} \mid \mu_{it}, \sigma_{it}^2) + \sum_{j \in \Omega_{it}} \log \left(\frac{1}{2} p(y_{ijt} \mid x_{ijt}, \theta_{it}, \mu_{jt} - \sigma_{jt}) + \frac{1}{2} p(y_{ijt} \mid x_{ijt}, \theta_{it}, \mu_{jt} + \sigma_{jt}) \right) \\ = & \quad c^* - \frac{(\theta_{it} - \mu_{it})^2}{2\sigma_{it}^2} + \sum_{j \in \Omega_{it}} \log \left(\frac{1}{2} p(y_{ijt} \mid x_{ijt}, \theta_{it}, \mu_{jt} - \sigma_{jt}) + \frac{1}{2} p(y_{ijt} \mid x_{ijt}, \theta_{it}, \mu_{jt} + \sigma_{jt}) \right), \end{aligned} \tag{4.8}$$

where c and c^* are constants not involving θ_{it}. To approximate (4.8) with the log of a normal density, we apply a one-step Newton-Raphson update centered at the prior mean to estimate the posterior mean, and approximate the posterior variance using the reciprocal of the second derivative of the log-posterior, also evaluated at the prior mean. This procedure requires evaluating the first and second derivatives of the log-posterior. Define

$$u_{ijt}(\theta_{it}) = \frac{1}{2} p(y_{ijt} \mid x_{ijt}, \theta_{it}, \mu_{jt} - \sigma_{jt}) + \frac{1}{2} p(y_{ijt} \mid x_{ijt}, \theta_{it}, \mu_{jt} + \sigma_{jt}).$$

Then the first and second derivatives of the log-posterior with respect to θ_{it} are

$$\frac{\partial \log p(\theta_{it} \mid \boldsymbol{D}_t)}{\partial \theta_{it}} = \frac{-(\theta_{it} - \mu_{it})}{\sigma_{it}^2} + \sum_{j \in \Omega_{it}} \frac{\frac{\partial}{\partial \theta_{it}} u_{ijt}(\theta_{it})}{u_{ijt}(\theta_{it})},$$

$$\frac{\partial^2 \log p(\theta_{it} \mid \boldsymbol{D}_t)}{\partial \theta_{it}^2} = \frac{-1}{\sigma_{it}^2} + \sum_{j \in \Omega_{it}} \left(\frac{u_{ijt}(\theta_{it}) \left(\frac{\partial^2}{\partial \theta_{it}^2} u_{ijt}(\theta_{it}) \right) - \left(\frac{\partial}{\partial \theta_{it}} u_{ijt}(\theta_{it}) \right)^2}{(u_{ijt}(\theta_{it}))^2} \right). \tag{4.9}$$

The one-step Newton-Raphson update for approximating the posterior mean is then given by

$$\mu_{it}^* = \mu_{it} - \frac{\frac{\partial \log p(\theta_{it} | \boldsymbol{D}_t)}{\partial \theta_{it}} \Big|_{\theta_{it}=\mu_{it}}}{\frac{\partial^2 \log p(\theta_{it} | \boldsymbol{D}_t)}{\partial \theta_{it}^2} \Big|_{\theta_{it}=\mu_{it}}}. \tag{4.10}$$

The corresponding approximation for the posterior variance is

$$\sigma_{it}^{*2} = \left(\frac{\partial^2 \log p(\theta_{it} \mid \boldsymbol{D}_t)}{\partial \theta_{it}^2} \Big|_{\theta_{it}=\mu_{it}} \right)^{-1}. \tag{4.11}$$

This formulation allows posterior updates to be computed independently for each player, making the algorithm highly parallelizable and well-suited to large-scale implementation. The computational details for evaluating (4.10) and (4.11) are provided in Appendix A, included in the supplementary materials. Appendix B, also included in the supplementary materials, then quantitatively assesses the accuracy of the approximation used in (4.6), comparing the resulting per-game posterior updates with high-precision Gauss–Hermite evaluations. While it may be preferable to evaluate (4.11) at $\theta_{it} = \mu_{it}^*$ instead of $\theta_{it} = \mu_{it}$, the expressions as written allow for computing μ_{it}^* and σ_{it}^* in parallel. This approach, evaluating the posterior variance at the prior mean, was also adopted in the development of the Glicko system (Glickman, 1999).

5 System Parameter Optimization via Predictive Likelihood

The algorithm described in Sections 4.1–4.3 assumes that the system parameters α_0, α_1, β_0, β_1 and τ^2 are known or have been estimated in advance. One strategy for estimating these parameters is to maximize one-step-ahead predictive accuracy. This involves fixing the system parameters at candidate values, applying the updating algorithm for all games through period t, and then evaluating predictive performance using game outcomes observed in period $t+1$.

Let L_{kt} denote the likelihood contribution of game k during period t, where player i faces player j and the observed outcomes is y_{ijt}. This quantity is defined as

$$\begin{aligned} L_{kt} &= L(\alpha_0, \alpha_1, \beta_0, \beta_1, \tau^2 \mid y_{kt}) \\ &= \int \int \Pr(Y_{ijt} = y_{kt} \mid x_{ijt}, \theta_{it}, \theta_{jt}, \alpha_0, \alpha_1, \beta_0, \beta_1) p(\theta_{it} \mid \boldsymbol{D}_{t-1}) p(\theta_{jt} \mid \boldsymbol{D}_{t-1}) d\theta_{it} d\theta_{jt}. \end{aligned} \tag{5.1}$$

Because the double integral in (5.1) does not admit a closed-form expression, we approximate it using 3-point Gauss-Hermite quadrature for each of the marginal distributions $p(\theta_{it} \mid \boldsymbol{D}_{t-1})$ and $p(\theta_{jt} \mid \boldsymbol{D}_{t-1})$. This results in a grid of 9 evaluation points over the joint distribution of $(\theta_{it}, \theta_{jt})$, corresponding to all pairwise combinations of three nodes from each marginal. We use a 3-point Gauss-Hermite quadrature here to ensure greater numerical precision in estimating predictive accuracy. This added accuracy is particularly important for model selection, where small improvements in fit may influence system parameter optimization. The integral is then approximated as a weighted sum of the likelihood values evaluated at these 9 pairs, with weights derived

from the product of the respective Gauss-Hermite weights. This quadrature-based approximation yields an efficient and accurate method for evaluating predictive likelihoods, enabling system parameter selection via out-of-sample validation.

Given a set of candidate system parameter values, the strength updating algorithm is first run through a designated training period $t_{\text{train}} < T$. For each subsequent time period $t = t_{\text{train}} + 1, \ldots, T$, the following steps are performed:

1. Compute the predictive log-likelihood for period t, denoted $\ell_t = \sum_k \log L_{kt}$, where the sum is over all games played in that period.
2. If $t = T$, stop. Otherwise, proceed to the next step.
3. Apply the strength update algorithm to the game outcomes from period t.
4. Increment t by 1 and return to step 1.

The cumulative predictive log-likelihood,

$$\sum_{t=t_{\text{train}}+1}^{T} \ell_t, \tag{5.2}$$

serves as the objective function for selecting the optimal system parameters.

Rather than relying on a fixed grid of candidate system parameter values, it is common practice to select them adaptively by directly optimizing the predictive log-likelihood. This can be done using gradient-based optimization methods or, alternatively, through derivative-free approaches such as the Nelder–Mead simplex algorithm (Nelder and Mead, 1965), which is available via the `optim` function in R. Given the potential multimodality of the objective function, it is advisable to run the optimization from multiple initial values to improve the chances of identifying a globally optimal solution.

Once the system parameters have been optimized and normal prior distributions for player strengths have been obtained, predictive probabilities for future games can be computed using the same approach to evaluate the game likelihood in (5.1). Specifically, for players i and j with prior strength distributions $\mathrm{N}(\theta_i \mid \mu_i, \sigma_i^2)$ and $\mathrm{N}(\theta_j \mid \mu_j, \sigma_j^2)$, we approximate $\Pr(Y_{ij} = y \mid x_{ij}, \mu_i, \mu_j, \sigma_i^2, \sigma_j^2)$, which integrates $\Pr(Y_{ijt} = y \mid x_{ij}, \theta_i, \theta_j)$ over the two normal priors, using Gauss-Hermite quadrature over a grid of nine evaluation points (or more, if higher accuracy is desired).

6 Application to ICCF Game Outcomes

We demonstrate the proposed updating algorithm using game outcomes from the ICCF, based on a dataset of games completed between January 2016 and March 2022. Only "normal" games were included. These are defined as games that were completed through play, excluding those resolved by default, adjudication, or exceeding time limits. The final dataset consisted of a total of 392,658 games involving 8,976 unique players. Rating periods were defined in 3-month intervals, consistent with the cadence used in the ICCF's previous Elo-based rating system. This partitioning resulted in 25 rating periods over the six-year window.

As part of the evaluation and implementation process, the ICCF was given the option to adopt a version of the rating system that incorporated a white-player advantage, represented by the parameters α_0 and α_1. In August 2022, the ICCF Congress approved the implementation of the new rating algorithm but voted against including a white advantage in the final system.

Table 1: Optimized values of system parameters.

Parameter	Optimized Value
β_0	0.35338
β_1	0.57041
τ	0.46040

This decision reflected concerns about the interpretability and fairness of incorporating a white-player advantage, despite its statistical justification. Accordingly, in the version of the algorithm described below, the parameters α_0 and α_1 are set to zero, reflecting the specification adopted by the ICCF.

6.1 System Parameter Tuning and Model Calibration

The system parameters β_0, β_1 and τ^2 were optimized using the computational procedure described in Section 5. Specifically, the first 20 time periods were treated as training data, while the remaining 5 periods were used to evaluate the predictive log-likelihood. The strength parameters were updated sequentially at each 3-month interval during this evaluation phase.

Prior distributions for player strengths at the start of the time series were based on available ICCF ratings. To facilitate this, we applied a transformation between ICCF Elo ratings and the model's latent strength parameters that preserved the ratio $\mathrm{P}(y_{ij} = 1)/\mathrm{P}(y_{ij} = 0)$ between players i and j. Specifically, for player i at time t, the ICCF rating R_{it} was linked to the model-based mean strength μ_{it} via

$$R_{it} = 1500 + \frac{400}{\log(10)}\mu_{it} = 1500 + 173.72 \cdot \mu_{it} \ . \tag{6.1}$$

For players with a rating in January 2016, we assumed a normal prior distribution with a mean μ_{it} converted from their rating based on (6.1) and a standard deviation of 100 Elo points to reflect moderate uncertainty. This corresponds to $\sigma_{it} = 100/173.72 = 0.576$ on the latent scale. For unrated players, we specified a default prior mean of 1800 Elo and standard deviation of 250, reflecting greater uncertainty in the absence of historical information. These values corresponded to a normal prior for θ_{it} with $\mu_{it} = 1.727$ and $\sigma_{it} = 1.439$. We explored alternative specifications for the prior distributions, but found that reasonable variations had minimal impact on the results.

The optimized values of the system parameters, determined by maximizing the predictive validity criterion described earlier, are reported in Table 1. The estimated standard deviation of the innovation process, τ, is approximately 0.460 on the logit scale (i.e., the latent scale on which outcome probabilities are modeled), corresponding to a rating volatility of about 80 Elo points over a 3-month period. This value reflects the typical degree of strength fluctuation permitted by the model between rating periods. The estimated value of β_0, which controls the baseline probability of a draw, corresponds to a draw rate of 0.416 between two evenly matched players rated at 1500 Elo. In contrast, the estimated β_1, which captures the increase in draw likelihood with player strength, implies a draw probability of 0.950 between two elite players each rated at 2500 Elo, underscoring the model's recognition that draws are nearly inevitable at the highest levels. Together, these estimates illustrate the model's capacity to flexibly capture the observed variation in draw frequencies across different skill levels.

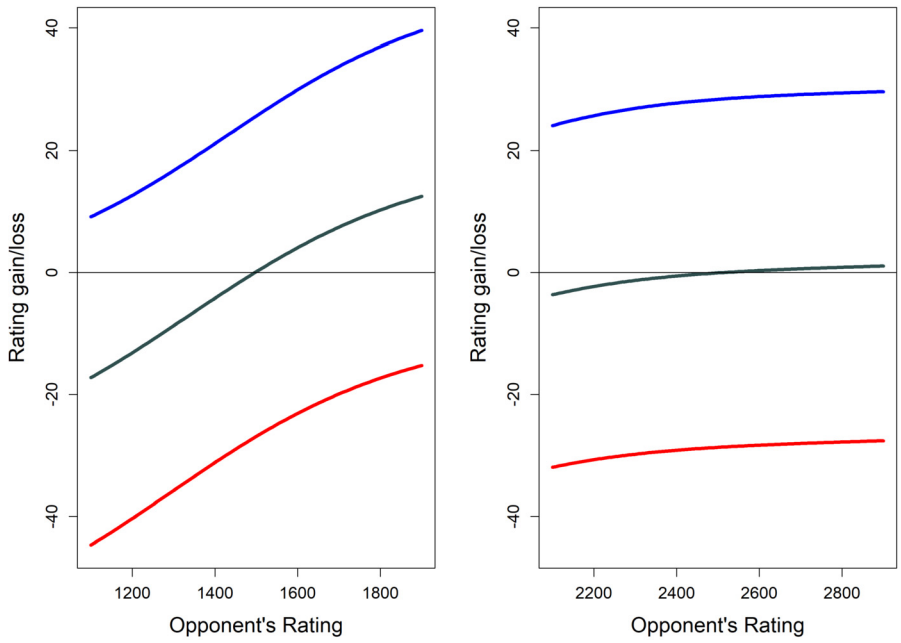

Figure 2: Each panel shows the rating change resulting from a win (blue), draw (green), or loss (red) against an opponent with a given Elo rating, based on the model parameters in Table 1. The left panel assumes that the focal player has a rating of 1500; the right panel assumes a rating of 2500. For all players, the prior standard deviation is assumed to be 100 Elo points.

The implications of the optimized parameter estimates are illustrated in Figure 2. The two panels show how rating changes vary with game outcomes and opponent strength under the fitted model. In the left panel, where the focal player is rated 1500 Elo, rating updates are highly sensitive to opponent rating: wins against stronger opponents lead to larger gains, while losses to weaker opponents incur greater drops. Draws result in modest rating increases when the opponent is stronger and modest losses when the opponent is weaker.

In the right panel, where the focal player is rated 2500, the magnitude of rating changes is much smaller across all outcomes. This reflects the model's tendency to treat results among strong players, especially draws, as expected, leading to minimal updates. In particular, draws yield near-zero rating changes across a wide range of opponent ratings, consistent with the model's strength-dependent draw probabilities. These patterns demonstrate how the model balances responsiveness to surprising results with stability in high-confidence settings.

6.2 Implementation Considerations and ICCF Decisions

While the system parameter values in Table 1 yielded optimal predictive accuracy, the resulting Elo-scale ratings were ultimately deemed unsuitable for practical use within the ICCF. According to feedback from ICCF officials, the ratings produced by the optimized system were unlikely to be accepted by the broader ICCF membership. This led to the adoption of a version of the model with modified system parameters, balancing predictive accuracy with the need for interpretability and practical acceptance.

The left panel of Figure 3 presents a scatter plot comparing the posterior mean ratings from the optimized model (converted to the Elo scale) with the official ICCF ratings as of

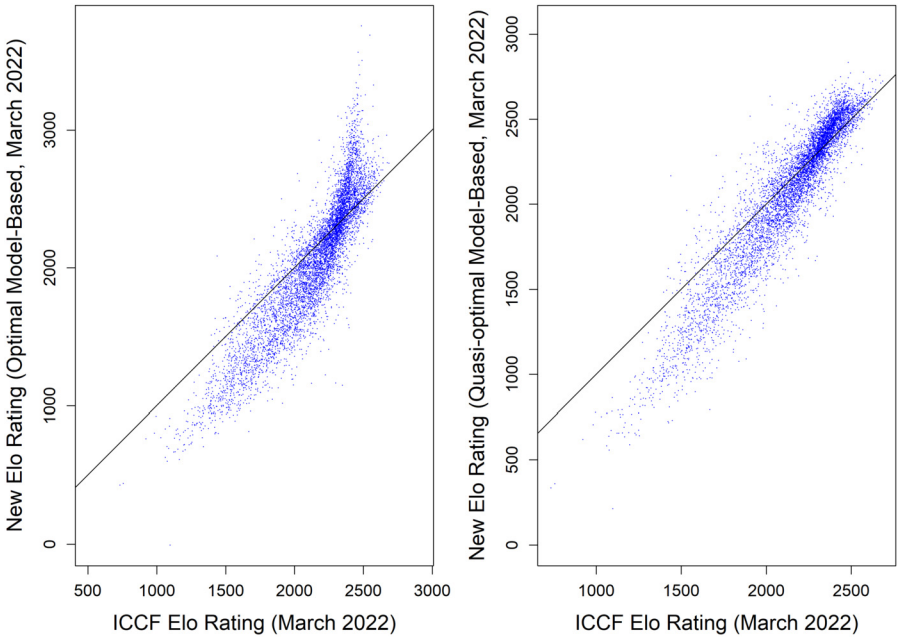

Figure 3: Comparison of new ratings in March 2022 versus official ICCF ratings under the Elo-based system. The ratings on the y-axis in the left panel are derived from the optimized system parameters in Table 1, and the ratings on the y-axis in the right panel are derived from the ICCF-implemented system parameters from Table 2. The line at $y = x$, indicating perfect agreement, is superimposed for reference.

March 2022. The diagonal line $y = x$ is included for reference. A key issue evident in the figure is that a substantial number of strong players with ICCF ratings of 2400 and above would have received ratings exceeding 3000 under the new system, as seen on the right side of the plot. Ratings of this magnitude are extremely rare in other Elo-based systems used for slow time-control chess and would likely lack credibility among ICCF members. The optimized value of $\tau = 0.460$, corresponding to an innovation standard deviation of approximately 80 Elo points, resulted in excessive volatility, substantially increasing variance from one period's posterior to the next period's prior. This level of fluctuation was deemed too large for a practical rating system.

Additionally, the optimized values of β_0 and β_1 were found to produce rating changes that were too small for drawn games between top players, undermining a primary motivation for revising the system, and too large for draws between average players. Together, these issues led to the conclusion that, despite strong predictive performance, the optimized system would not be suitable for implementation in its current form.

The rating system ultimately implemented by the ICCF used system parameter values shown in Table 2. In contrast to the values in Table 1, where β_0 and β_1 implied a draw probability of 0.416 between two players with strength parameters of 0 (i.e., Elo ratings of 1500), and 0.950 between two players with strength parameters of 5.756 (i.e., Elo ratings of 2500), we manually selected and then evaluated the system parameter values in Table 2. These were chosen to yield more moderate draw probabilities, corresponding to 0.6 and 0.8 for those same player strengths for players both rated 1500 and both rated 2500, respectively. These values produce larger rating

Table 2: ICCF-implemented values of model system parameters used in the updated ICCF rating system.

Parameter	ICCF-implemented value
β_0	1.09861
β_1	0.17037
τ	0.14391

changes for top players following drawn games and smaller changes for average players, which better align with ICCF priorities. Additionally, the lower value of $\tau = 0.14391$, corresponding to an innovation standard deviation of 25 Elo points, was chosen to ensure more modest rating fluctuations across time periods, leading to a more stable and interpretable rating system.

While the cross-entropy, defined as the negative predictive log-likelihood on a per-game basis (that is, the negative of the value in (5.2) averaged over games rather than summed), for the optimized parameters was 0.5063, the corresponding value for the ICCF-implemented parameters was 0.5724. Among the 17,414 games in the validation sample, 29.6% were decisive and 70.4% were draws. If predictions were made at random according to these empirical outcome frequencies, assigning equal probabilities to wins and losses within the decisive outcomes, the expected cross-entropy would be

$$-\big[(0.296/2)\log(0.296/2) + (0.296/2)\log(0.296/2) + 0.704\log(0.704)\big] = 0.8126.$$

Both fitted systems therefore perform substantially better than this random-prediction baseline. The optimized system achieves a 37.7% reduction in cross-entropy relative to the baseline, while the ICCF-implemented system achieves a 29.6% reduction.

The effects of using the ICCF-implemented parameters are illustrated in the right panel of Figure 3, which compares the new ratings (as of March 2022) to the ICCF's legacy Elo-based ratings. The figure shows that the ratings for top players remain below 3000, improving face validity relative to the fully optimized system. While lower-rated players tend to receive lower ratings under the new system than under the legacy ICCF system, these outcomes were judged to be consistent with expectations and acceptable for implementation.

An additional refinement in the implemented system was a cap on the growth of a player's prior standard deviation from one period to the next. Specifically, once a player's uncertainty exceeded a certain threshold, it was no longer allowed to increase further. The rationale for this rule is that once a player has competed enough to obtain a reasonably reliable strength estimate, their uncertainty should not continue to grow indefinitely simply due to inactivity. In practice, the ICCF system was configured to apply the usual innovation variance τ^2 to a player's posterior variance when computing the prior for the next period only if the player's posterior standard deviation was below 0.691 (equivalent to 120 Elo points). If the posterior standard deviation was already at or above this threshold, it was carried forward unchanged as the prior standard deviation in the next time period.

The ICCF-implemented system parameter values in Table 2 yield ratings with strong predictive validity, as demonstrated in Figure 4. The analysis is based on the five validation periods of ICCF game results. The left panel displays a side-by-side boxplot of pre-game predicted draw probabilities for all 17,414 games in the validation sample. The left box corresponds to games that ended decisively, while the right box corresponds to games that ended in a draw. As expected from a well-calibrated model, draw probabilities tend to be higher for games that were actually

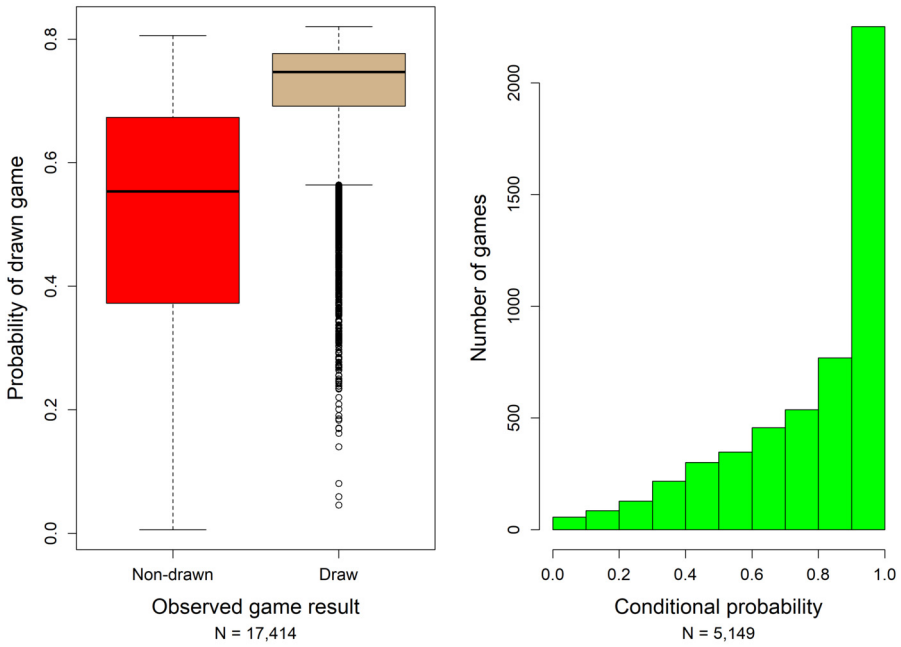

Figure 4: Left panel: boxplots of predicted draw probabilities for the 17,414 games in the validation sample, separated by whether the game was decisive or drawn. Right panel: histogram of predicted probabilities of the observed outcome (conditional on a decisive result) for the 5,149 decisive games in the validation set. All predictions are based on the ICCF-implemented system parameters in Table 2.

drawn, supporting the validity of the system's draw predictions. The right panel focuses on the subset of 5,149 games from the validation set that resulted in a decisive outcome. It shows a histogram of the predicted probabilities assigned to the observed (winning) outcome, conditional on the game not being a draw. Specifically, these values take the form $p = p_{win}/(p_{win} + p_{loss})$, omitting the draw probability. The distribution skews toward higher probabilities, indicating that the model generally assigns higher likelihood to outcomes that actually occurred, another sign of predictive validity. Moreover, only 14.8% of decisive outcomes were assigned probabilities below 0.5, indicating that true upsets were relatively rare. Together, these analyses provide empirical support for the effectiveness of the developed rating system and its ability to generate credible and informative predictions.

7 Discussion

This paper introduced a dynamic rating system for games with possible outcomes of a win, loss, or draw, where the probability of a draw is modeled as a function of player strength. Building on the model developed by Glickman (2025), the proposed method incorporates a state-space framework in which player abilities evolve over time according to a normal random walk. The resulting system enables real-time updating of ratings using an approximate Bayesian filtering algorithm, ensuring scalability and responsiveness to new results. This model was applied to game data from the ICCF, leading to the development and implementation of a new rating system that addresses long-standing issues in handling drawn results, particularly among top-level players.

Although the present analysis focuses on chess, the underlying modeling framework extends naturally to other competitive settings in which head-to-head games may result in ties. Team sports such as soccer share this structure, and in such contexts, Elo-type or Bradley–Terry formulations are widely used to infer team strengths. Incorporating a strength-dependent tie component, as in the model developed here, would allow such systems to capture the empirical tendency for ties to occur more frequently between evenly matched teams. This phenomenon was discussed more fully by Glickman (2025). Related statistical approaches have been developed for soccer and similar sports using bivariate count models that explicitly represent correlated team scores, such as the bivariate Poisson and diagonal-inflated bivariate Poisson formulations of Karlis and Ntzoufras (2003, 2005), and the more recent Bayesian bivariate Conway–Maxwell–Poisson regression model of Florez et al. (2025). While these models treat goals directly rather than win-tie-loss outcomes, their treatment of dependence and tie frequency is conceptually similar to the structure of the present rating framework, suggesting that analogous extensions could be developed for rating systems in games and sports where ties are non-negligible.

While the implemented chess rating system represents a significant step forward, there are limitations and opportunities for further refinement. Although the ICCF ultimately chose not to implement the system with an explicit white advantage, the modeling framework is fully capable of incorporating such asymmetries. Additionally, the current method assumes a linear relationship between average player strength and both the probability of a draw and the advantage of playing white. While these forms are convenient and interpretable, another possible extension is to introduce nonparametric modeling for draw probabilities and white advantage, such as splines or Gaussian processes, to better capture complex dependencies observed in real-world data. However, such flexibility must be balanced against the need for computational efficiency in real-time rating environments.

With its adoption of this system, the ICCF became the first major competitive organization to implement a rating algorithm that explicitly models the probability of a draw as a distinct, strength-dependent outcome. This innovation represents an important evolution in rating methodology, particularly in domains where drawn outcomes are prevalent and carry nuanced information about player ability. The ICCF approved the system in 2022, and it has been in operational use since 2023, providing players and officials with a more accurate and context-aware evaluation of performance.

Supplementary Material

- `Appendices.pdf`: Appendices A and B.
- `Code_and_Data.zip`: Zip file consisting of code and data to run the analyses in this manuscript.

Acknowledgement

Thanks to Austin Lockwood and the International Correspondence Chess Federation for providing the data used in this study, and for helpful feedback.

Funding

This work was partially supported by the International Correspondence Chess Federation.

References

Bhat CR (1995). A heteroscedastic extreme value model of intercity travel mode choice. *Transportation Research. Part B: Methodological*, 29(6): 471–483. https://doi.org/10.1016/0191-2615(95)00015-6

Boys R, Dunsmore I (1987). Diagnostic and sampling models in screening. *Biometrika*, 74(2): 365–374. https://doi.org/10.1093/biomet/74.2.365

Bradley RA, Terry ME (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4): 324–345. https://doi.org/10.2307/2334029

Crouch EA, Spiegelman D (1990). The evaluation of integrals of the form $\int_{-\infty}^{\infty} f(t)\exp(-t^2)$: Application to logistic-normal models. *Journal of the American Statistical Association*, 85(410): 464–469. https://doi.org/10.1080/01621459.1990.10476222

David H (1988). *The Method of Paired Comparisons*. Charles Griffin & Company, London.

Davidson RR (1970). On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329): 317–328. https://doi.org/10.1080/01621459.1970.10481082

Davidson RR, Beaver RJ (1977). On extending the Bradley-Terry model to incorporate within-pair order effects. *Biometrics*, 33(4): 693–702. https://doi.org/10.2307/2529467

Durbin J, Koopman SJ (2012). *Time Series Analysis by State Space Methods*, volume 38. OUP, Oxford.

Elo AE (1978). *The Rating of Chess Players, Past and Present*. Arco Publishing, New York.

Fahrmeir L, Tutz G (1994). Dynamic stochastic models for time-dependent ordered paired comparison systems. *Journal of the American Statistical Association*, 89(428): 1438–1449. https://doi.org/10.1080/01621459.1994.10476882

Florez M, Guindani M, Vannucci M (2025). Bayesian bivariate Conway–Maxwell–Poisson regression model for correlated count data in sports. *Journal of Quantitative Analysis in Sports*, 21(1): 51–71. https://doi.org/10.1515/jqas-2024-0072

Glickman ME (1993). Paired comparison models with time-varying parameters, Ph.D. thesis, Harvard University.

Glickman ME (1995). A comprehensive guide to chess ratings. *American Chess Journal*, 3(1): 59–102.

Glickman ME (1999). Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, 48(3): 377–394. https://doi.org/10.1111/1467-9876.00159

Glickman ME (2001). Dynamic paired comparison models with stochastic variances. *Journal of Applied Statistics*, 28(6): 673–689. https://doi.org/10.1080/02664760120059219

Glickman ME (2025). Paired comparison models with strength-dependent ties and order effects. Statistical Modelling. In press. https://doi.org/10.1177/1471082X251400474.

Glickman ME, Jones AC (2024). Models and rating systems for head-to-head competition. *Annual Review of Statistics and Its Application*, 12: 259–282. https://doi.org/10.1146/annurev-statistics-040722-061813

Gorgi P, Koopman SJ, Lit R (2019). The analysis and forecasting of tennis matches by using a high dimensional dynamic model. *Journal of the Royal Statistical Society. Series A. Statistics in Society*, 182(4): 1393–1409. https://doi.org/10.1111/rssa.12464

Harding MC, Hausman J (2007). Using a Laplace approximation to estimate the random coefficients logit model by nonlinear least squares. *International Economic Review*, 48(4): 1311–1328. https://doi.org/10.1111/j.1468-2354.2007.00463.x

Hastie T, Tibshirani R (1986). Generalized additive models. *Statistical Science*, 1(3): 297–310. https://doi.org/10.1214/ss/1177013604

Ingram M (2021). How to extend Elo: A Bayesian perspective. *Journal of Quantitative Analysis in Sports*, 17(3): 203–219. https://doi.org/10.1515/jqas-2020-0066

Karlis D, Ntzoufras I (2003). Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society. Series D. The Statistician*, 52(3): 381–393.

Karlis D, Ntzoufras I (2005). Bivariate Poisson and diagonal inflated bivariate Poisson regression models in R. *Journal of Statistical Software*, 14: 1–36. https://doi.org/10.18637/jss.v014.i10

Knorr-Held L (2000). Dynamic rating of sports teams. *Journal of the Royal Statistical Society. Series D. The Statistician*, 49(2): 261–276.

Nelder JA, Mead R (1965). A simplex method for function minimization. *The Computer Journal*, 7(4): 308–313. https://doi.org/10.1093/comjnl/7.4.308

Pirjol D (2013). The logistic-normal integral and its generalizations. *Journal of Computational and Applied Mathematics*, 237(1): 460–469. https://doi.org/10.1016/j.cam.2012.06.016

Pryanishnikov I, Zigova K (2016). Multinomial logit models for the Austrian labor market. *Austrian Journal of Statistics*, 32(4): 267–282. https://doi.org/10.17713/ajs.v32i4.461

Steen N, Byrne G, Gelbard E (1969). Gaussian quadratures for the integrals $\int_0^\infty e^{-x^2} f(x)dx$ and $\int_0^b e^{-x^2} f(x)dx$. *Mathematics of Computation*, 23(107): 661–671. https://doi.org/10.1090/S0025-5718-1969-0247744-3

Szczecinski L, Djebbi A (2020). Understanding draws in Elo rating algorithm. *Journal of Quantitative Analysis in Sports*, 16(3): 211–220. https://doi.org/10.1515/jqas-2019-0102

West M, Harrison PJ, Migon HS (1985). Dynamic generalized linear models and Bayesian forecasting. *Journal of the American Statistical Association*, 80(389): 73–83. https://doi.org/10.1080/01621459.1985.10477131

Wood SN (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1): 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x